Proceedings of the Institute of Acoustics

A TWO DIMENSIONAL KINEMATIC MAPPING BETWEEN SPEECH ACOUSTICS AND VOCAL TRACK CONFIGURATIONS

A. Hatzis Dept. of Computer Science, University of Sheffield, Sheffield S1 4DT
a.hatzis@dcs.shef.ac.uk, www.dcs.shef.ac.uk/~nassos
P.D. Green Dept. of Computer Science, University of Sheffield, Sheffield S1 4DT
p.green@dcs.shef.ac.uk, www.dcs.shef.ac.uk/~pdg

1. INTRODUCTION

The traditional "articulatory" phonetic theory of Melville Bell, **[3]**, supported the view that the phonetic quality of vowels is derived from the position and height of the point of constriction of the tongue. "*The most successful outcome of this idea, and one still in use for vowel description, is the set of 'cardinal vowels' devised by Daniel Jones*", **[5]**. The typical IPA quadrilateral diagram shows the relative vowel qualities for the English language according to hypothetical positions of the tongue for front-back and high-low, dimensions. Ladefoged correctly observes, **[7]**, that phoneticians may have been using these articulatory descriptions as labels to specify acoustic dimensions rather than as descriptions of actual tongue positions. This is confirmed by the construction of the vowel quadrilateral. Although Jones started his cardinal vowel system by defining articulatory configurations for two of the cardinal vowels /i/ and /ɑ/, the positions of the other six cardinal vowels were based on these two and defined according to the number of auditorily equidistant steps between /i/ and /ɑ/, **[7]**. Therefore the cardinal vowel system is best described by auditory qualities rather than articulatory specifications. Most of the cardinal vowels, often called pure vowels because of their single stable auditory quality, have not been adequately described in terms of their corresponding stable articulatory configuration. These cardinal vowel target configurations are the basis for the vowel maps described in this paper. Similarly front-back and high-low directions are not fully related to articulatory movements of the tongue. Jones implies that the specifications of tongue position suggest an invariant tongue position for each vowel quality. This is not true : several experimental investigations, **[8]**, **[9]**, showed that speakers can produce vowels with the same auditory quality using compensatory articulation, e.g. jaw aperture, larynx height, lip radiation. This important limitation is mostly evident because of the modern method used to construct vowel charts.

Nowadays, vowel charts frequently used in speech training packages, are based on formant frequency, the front-back direction corresponds to F2-F1 and the high-low direction corresponds to F1. There are several arguments favouring the use of whole spectra, **[1]**, instead of formants extraction. First, the algorithms designed to estimate the short-term spectral envelope are relatively fast compared with the speed of the more complex algorithms needed in formant detection. Second, not all the phoneme categories can be represented sufficiently by formant patterns, as Rabiner and Schafer observe, **[10]**, *"in most consonants the first formant is either not observable or at a very low frequency. The frequency of the first formant will be at a minimum in most consonants in*

which there is an articulatory closure, but the frequency of the second formant varies considerably". Third the formants of misarticulated sounds are often poorly defined, and fourth the two-formant plot does not usually account for F3, which also contributes to vowel quality, **[2]**. For these reasons, contrast between consonantal production, and secondary articulation is not feasible.

We propose here an alternative method which we call optico-acoustic articulography (OPTACIA), which maps not only vowels but any speech sound from acoustics to 2D articulatory positions. In this paper we demonstrate the method by attempting to construct an extended version of a "quadrilateral" vowel map.

2. Map Construction

2.1 Acoustic vector representation

In capturing the spectral envelope for the speech it is desirable to reduce the effects of pitch in order to focus on the time-varying properties of the articulators. We therefore chose cepstral analysis to deconvolve the vocal tract filter. OLTK, a software developed by the authors, **[6]**, uses mel-frequency cepstral coefficients as its acoustic vector representation. Eight coefficients together with overall energy are used. The OPTACIA map was created by first defining the positions of articulators for the minimal set of reference sounds, 'cardinals', needed to construct the map **(§2.3)**. This is the relationship between the acoustics and articulation used in the sort of transformation we imply here. The acoustics were produced by pronouncing these sounds in isolation and prolonging them for 30secs. Then they were subjected to cepstral analysis, and the resulting labelled 9-component vector formed the training input to a multi-layer perceptron (MLP), which has two outputs, corresponding to a fixed X and Y positions on the map chosen for each one of the articulatory gestures of the cardinals.

2.2 Artificial neural network (ANN) mapping

If we consider the ANN as a mapping function from an input space, the 9D cepstral representation, to an output space, the 2D coordinates of a specific articulatory configuration, then if every part of its input space that has been trained with sufficient representative acoustic vectors the ANN should behave consistently. In practice this means that a small diffrenence in the acoustics corresponding to a specific gesture will result in mapping onto areas close to the fixed points defined for the 'cardinals'. Two questions arise here: how similar the acoustics of speech production have to be to those used for training the map, and what kind of mapping we get for acoustics that correspond to vocal tract configurations that differ significantly from those used in the training stage. Two important observations are made here by Bishop, **[4]**: first, there could be a lower dimensional space where our data points can be restricted, assuming that features are generally correlated in some way; and, second, we hope that *"the value of the output variables will not change arbitrarily from one region of input space to another, but will typically vary smoothly as a function of the input variables. Thus, it is possible to infer the values of the output variables at intermediate points, where no data is available, by a process similar to interpolation".* We comment on these effects in the next paragraph.

2.3 The OPTACIA cardinal maps

2.3.1 Articulatory definition on the minimum set of our cardinal vowels

We mentioned in the introduction, that Jones started his cardinal chart by defining articulatory configurations for the two vowels that appear to be at the limits of articulation for the tongue-jaw combination: /i/ most high and front and /á/ most low and back. We similarly started by fixing positions in the map for the gestures of these two vowels. *"/i/ is produced with the tongue as high and as far forward in the mouth as it is possible to go without causing audible friction. /á/ is produced with the tongue as low and retracted as possible"*, **[5]**. With only these two vowels the ANN does not have enough information to generalise the output for the ceptstral represenation of the other cardinals in the vowel space. For example the /ae/ sound would be plotted close to the area of the /á/ sound. Therefore we needed to define and train the network to distinguish and learn a more accurate mapping between the positions of a minimum of four cardinal vowels, (**Figure 1**). So we trained the ANN for the /ae/ sound, produced with the jaw as down as possible and the tongue blade at a slightly raised position or the tip slighty retracted, (otherwise the sound quality is that of an /å). Finally for /u/ the jaw is at the same position as that of /i/ but the tongue is retracted. It is important to mention here that we were careful to ensure that productions of these four cardinal vowels were made without any lip protrusion and rounding or lip spreading. However, another parameter, the "width" of the vowel, which is related to the movement of the root of the tongue was active during our productions. For example /i/ was pronounced with the root of the tongue being advanced, drawn forward, and widening the larynx.

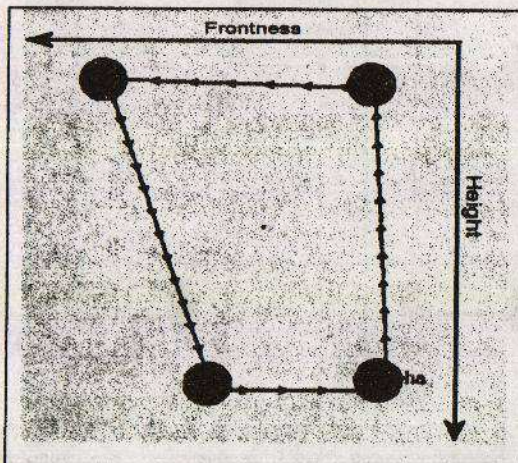

Figure 1: Fixed mapping of four cardinal vowels, /i/, /ae/, /á/, /u/

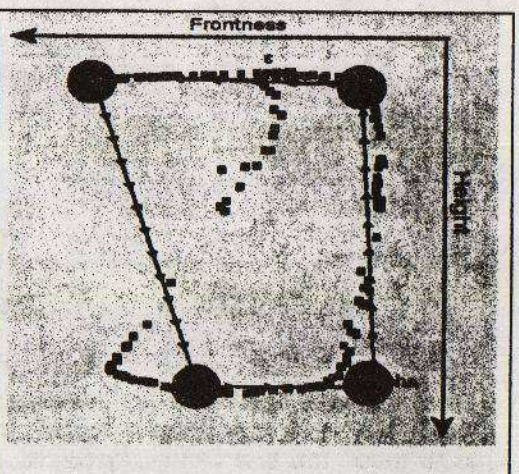

Figure 2 : Gliding from /i/ to /ae/ to /á/ to /u/. Gliding from /i/ to /ae/ passes through the junction of /å sound. Dark squares indicate the frame-by-frame acoustics mapping

2.3.2 Lingual mapping

Once we built the map we made the "gliding test". This is simply a test where the speaker is allowed to vary only the tongue and jaw in order to glide from /i/ to /ae/ to /á/ to /u/ and back to /i/, **(Figure 2)**. This way we test the vowel height and vowel frontness and see how they corellate with the movement of the tongue and jaw. As we can see in **Figure 2** the only direction of gliding where the map failed to generalise was that of /i/ to /ae/, where passing through /å quality causes the network to divert mapping and plot the sound close to the /u/ area of the map. All other directions were plotted consistently. In response, the next step was to create another map with five vowels this time, by fixing a position of the cardinal vowel /å, **(Figure 3)**. Ôhis improved the plotting direction of glide from /i/ to /ae/, passing from the intermediate position of the /å sound. According to phoneticians this is also what one expects for the dipthong /ai/, **(Figure 4)**. Because of the vowel space continuum, we get a transition that passes from the articulatory area of the /å sound. Mapping is also consistent when plotting the acoustics of the articulatory target of /o/, despite the fact it has not been trained with samples of this vowel. This is plotted close to the mapped area of the /u/ target across the direction of the glide between /á/ and /u/, **(Figure 5)**.

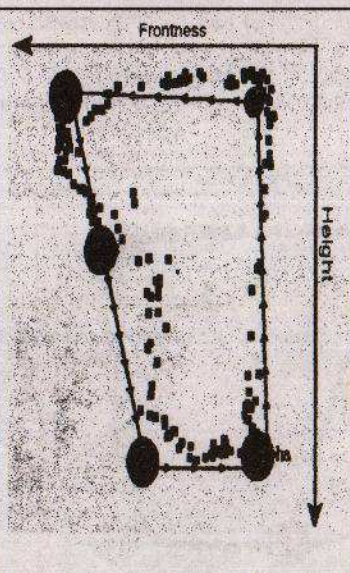

Figure 3 : Gliding from /i/ to /ae/ to /á/ to /u/. This time with the additional target of /å sound.

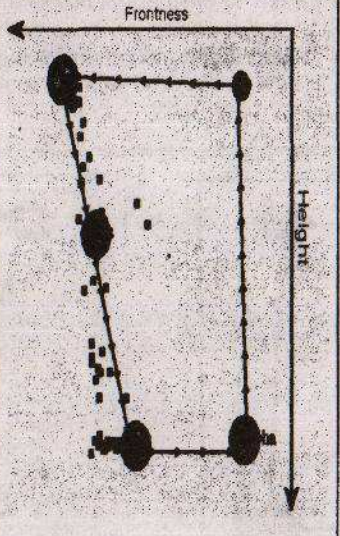

Figure 4 : Testing the production of the dipthong /ai/

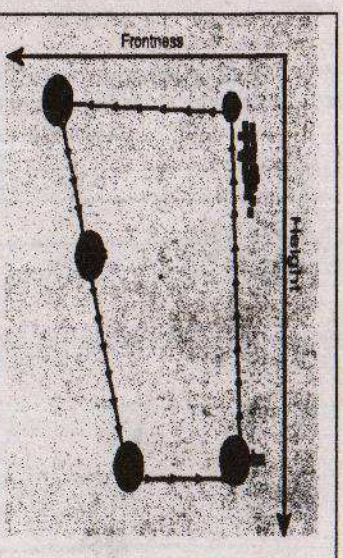

Figure 5 : Production of the /o/ sound and the relevant mapping.

3. The mapping effect of secondary articulation

3.1 Labialisation

In general, labialisation has the effect of plotting sounds near the areas of the cardinal vowels. In the case of /á/ the rounding of the lips produced acoustics similar to those for an /o/ or /u/ sound and

the mapping was along the direction of /á/ - /u/, **(Figure 8)**. On the other hand lip spreading on the same vowel caused the ANN to map the acoustics along the /á/ - /ae/ direction, **(Figure 9)**. Similarly, rounding the lips while producing the /i/ sound with the jaw lowered a little and the tongue slightly retracted produces the /y/ vowel whose aural quality is close to that of the /u/, **(Figure 7)**. Finally applying lip spreading on the /ae/ configuration gave a different quality to that sound but mapped relatively close to the /ae/ area, **(Figure 6)**.

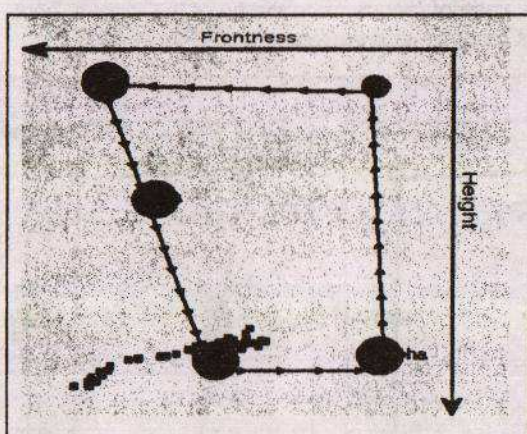

Figure 6 : The vowel /ae/ produced with lip spreading.

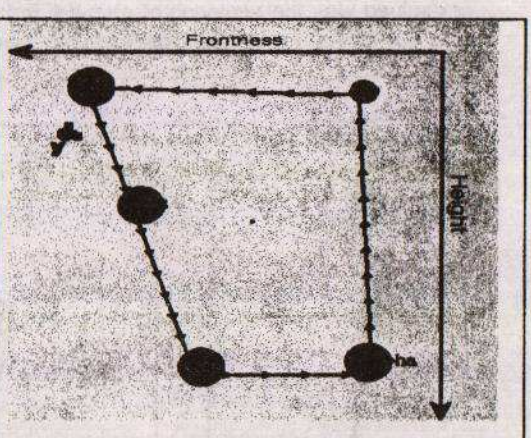

Figure 7 : Visual contrast between the productions of /i/ and /y/

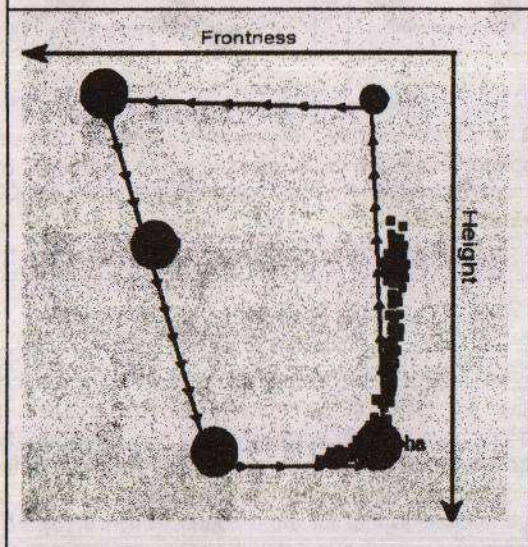

Figure 8 : The effect of lip rounding on the /á/ production.

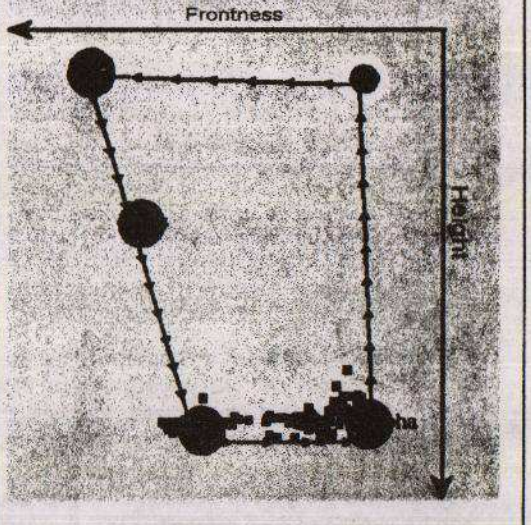

Figure 9 : The effect of lip spreading on the /á/ production.

3.2 Nasalisation – Rhotacisation – Pharyngealisation

Two more examples on the mapping effect of secondary articulation are the nasalised and pharyngealised versions of the /i/ sound. The first was produced by keeping the same lingual-jaw configuration as /i/ but raising the velum and the second by narrowing the pharynx with the root of the tongue. While nasalisation mapped the sound close to the area of /i/, **(Figure 10)**, pharyngelisation mapped it further away and closer to the /å area, **(Figure 11)**. Another quality we examined was the r-coloring of a vowel. In the case of /ae/, rainsing the tip of the tongue, gave us this quality. As a result the r-colored sound was mapped near the /ae/ sound, **(Figure 12)**.

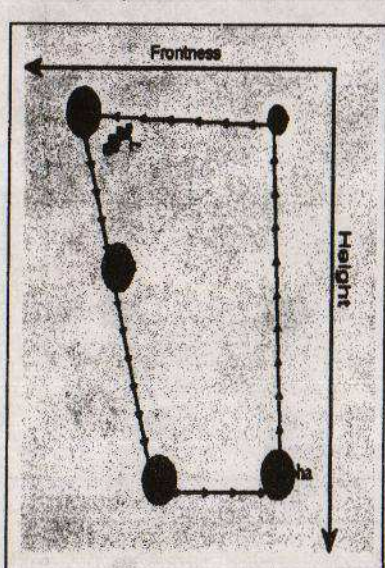

Figure 10 : Visual contrast between normal production of an /i/ sound and a nasalised one /i~/

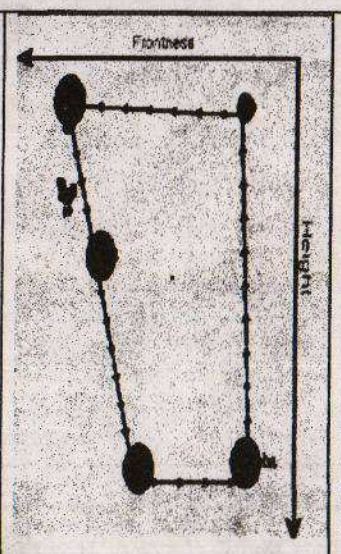

Figure 11 : Visual contrast between normal production of an /i/ sound and a pharyngealised one /i+/

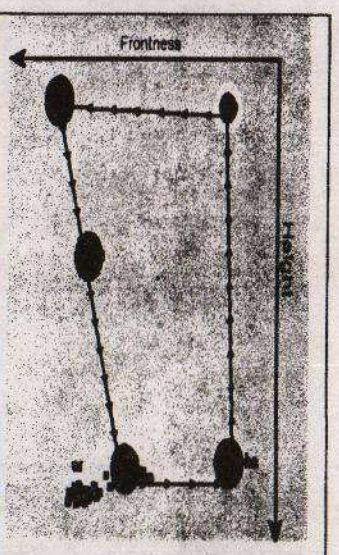

Figure 12 : Visual contrast between normal production of an /ae/ sound and the r-colored version of it /ar/.

3.3 Consonantal production

It is interesting to see what happens if we try to produce a consonant on the vowels map. We took first as an example the voiced sibilant fricative /z/. Although perceptually /i/ and /z/ are very dissimilar, there is a close match between the articulatory configurations of the two sounds. The main articulatory difference is the degree of stricture between the tongue and the palate and in the case of /z/ the tongue is raised a little and retracted from the /i/ configuration to produce the /z/ sound. Thus the plotted position of the /z/ sound is close to that for an /i/ sound, **(Figure 13)**. This can be explained if we remember that the ANN tries to generalise and find the closest match to what has already been trained with. This can be further justified by studying the plotting for the /z-ae/ syllable, a similar case to the production of the /ai/ dipthong. There is again the transition that passes from the area of the /å sound and the plotting along the direction of /i/-/ae/.

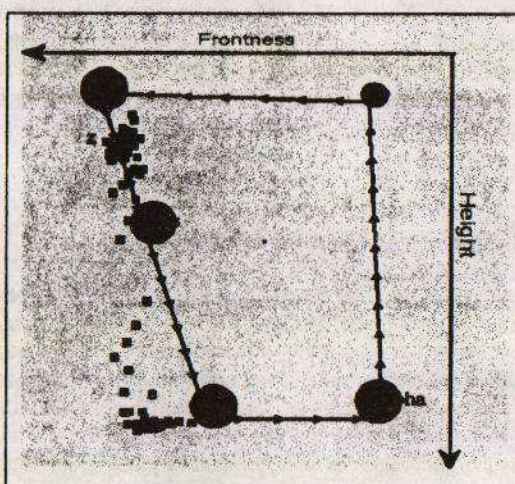

Figure 13 : Mapping of the /z/ sound and the transition to the /ae/ sound.

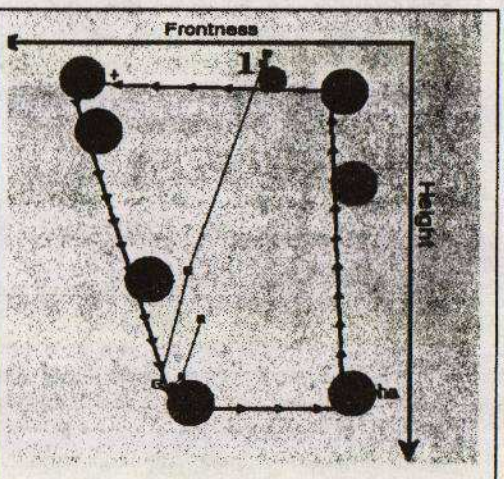

Figure 14 : Mapping of the /l/ sound and the transition to the /ae/ sound.

A more detailed mapping

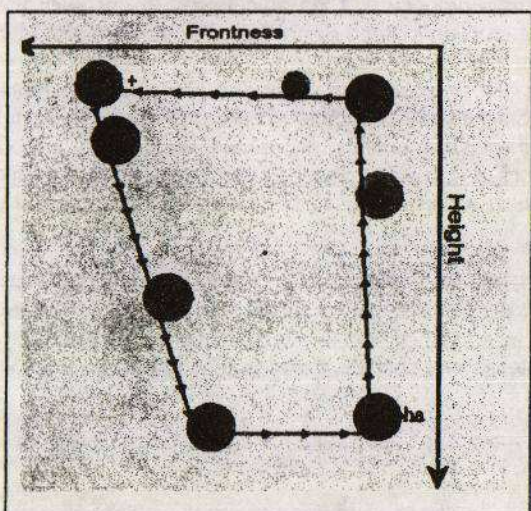

Figure 15 : A cardinal map with ten vowels

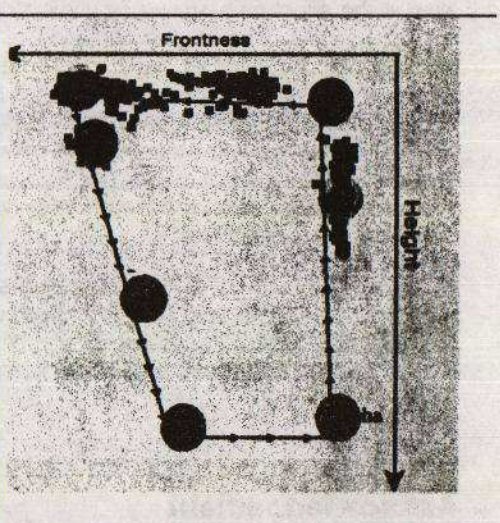

Figure 16 : The testing of the /i~/, /i+/, /y/ and /o/ sounds on a more detailed cardinal map

In order to define and visualise more closely the relationship between the acoustics and various articulatory configurations we designed another map, **(Figure 15)**, with ten targets. We used two extra targets to map the nasalised, pharyngealised /i/ at a position very close to the /i/ and three more targets for the articulatory configurations of /I/, /y/ and /o/. A testing of this more detailed map can be seen in **Figure 16**. The production of syllable /I-ae/ was also tested in the same map and transition was plotted, **(Figure 14)**.

4. Envisaging the future of optico-acoustic articulography

There has always been a major difficulty in relating the auditory qualities of the sounds we produce with the movement of the articulators. This is mainly due to the fact that instrumentation to achieve this is expensive, complicated, and invasive, e.g. electropalatography, electomagnetic articulography and others. Optacia offers an attractive alternative by simply relying on capturing the acoustics of speech production and using the traditional source-filter model to relate them with articulation. Although the two-dimensional space is proved to be inadequate to map the acoustics from a much higher dimensional space, the simple and innovative technique of fixing articulatory configurations on a customised map and using an artificial neural network to generalise and learn the space between them seems to make the necessary step to visualise the kinematics of certain articulators while the acoustics of one phone blend into those of the next. It certainly cuts out many of the limitations that appear on the traditional vowel chart. Most importantly because of the way optacia is designed, each sound can be adequately described by defining the positions of articulators. In this way we make an effort to move away from the traditional abstract notion of the phoneme target with auditory qualities to that of an articulatory target with an articulatory description. Then we try to see what is the effect on mapping by varying certain properties of the articulators. In this way, these targets work like poles of attraction and it is interesting to see how far from the target the produced sound is mapped, whether the acoustics have radically changed or not, and whether the shape of the vocal tract has changed significantly. It is also interesting to study and see what properties of the articulators we vary, to glide from one configuration to another. There has not been so far a systematic analysis and description on how the distinct sounds, phonemes, we produce can vary from a reference point by moving our articulators, and how the gliding to some other distinct sound is achieved. This includes also the limits on which a certain articulatory target can change and still consider it as a variant form.

There is much research to be done in this specific area. Optacia is only the beginning, the method needs to be better defined and described but the authors believe that we have already reached a stage that the community can benefit whether as a teaching tool for phonetics, or as a training tool in speech training, or perhaps even as a tool to test some phonetic theories. The future will show what of these can be true.

Acknowledgement

A. Hatzis is funded by the NHS programme New and Emerging Applications of Technology (NEAT).

REFERENCES

1. Arends N. (1993), *"The visual speech apparatus. An aid for the speech training"*, Manuscript, Instituut voor Doven, Nijmegen, pp. 135-156.

2. Arends N., Povel D.J., Michielsen S., Claassen J., Feiter I. (1991), *"An Evaluation of the Visual Speech Apparatus (VSA)"*, Speech Communication (1991), 10, pp. 405 – 414

3. Bell, A. M. (1867) "Visible speech or self-interpreting physiological letters for the writing of all languages in one alphabet. London: Simpkin and Marshall

4. Bishop M., (1999), *"Latent variable models"*, in Learning in Graphical Models, Jordan M. (Ed.), NATO Science Series, MIT Press, Boston, pp. 372-402

5. Clark, J. and Yallop, C. (1991), *"An Introduction to Phonetics and Phonology"*, Oxford, Blackwell Inc.

6. Hatzis A., Green P.D., and Howard S. (1999), *"Visual Displays in Practical Auditory Phonetics Teaching"*, Phonetics Teaching & Learning Conference (PTLC'99), University College of London, U.K.

7. Ladefoged, P. (1967) "Three areas of experimental phonetics". London: Oxford University Press.

8. Lindau M. (1978), *"Vowel features"*, Language 54: 541-63

9. Lindblom B., and Ohman S. (eds) (1979), *"Frontiers of speech communication research"*. New York : Academic Press.

10. Rabiner L.R. and Schafer R.W. (1988), *"Digital Processing of Speech Signal"*, Prentice-Hall